Philos Stud (2019) 176:2243–2267
https://doi.org/10.1007/s11098-018-1124-5

Rational social and political polarization

Daniel J. Singer[1] · **Aaron Bramson**[2,3,4] · **Patrick Grim**[5,6] · **Bennett Holman**[7] · **Jiin Jung**[8] · **Karen Kovaka**[9] · **Anika Ranginani**[1] · **William J. Berger**[1]

Published online: 13 June 2018

Abstract Public discussions of political and social issues are often characterized by deep and persistent polarization. In social psychology, it's standard to treat belief polarization as the product of epistemic irrationality. In contrast, we argue that the persistent disagreement that grounds political and social polarization can be produced by epistemically rational agents, when those agents have limited cognitive resources. Using an agent-based model of group deliberation, we show that groups of deliberating agents using coherence-based strategies for managing their limited resources tend to polarize into different subgroups. We argue that using that strategy is epistemically rational for limited agents. So even though group polarization looks like it must be the product of human irrationality, polarization can be the result of fully rational deliberation with natural human limitations.

Keywords Polarization · Epistemic rationality · Group deliberation · Social epistemology

✉ Daniel J. Singer
singerd@phil.upenn.edu

[1] University of Pennsylvania, Philadelphia, PA, USA

[2] Riken Brain Science Institute, Wakoshi, Japan

[3] Ghent University, Ghent, Belgium

[4] University of North Carolina at Charlotte, Charlotte, NC, USA

[5] University of Michigan, Ann Arbor, MI, USA

[6] Stony Brook University, Stony Brook, NY, USA

[7] Underwood International College, Yonsei University, Seoul, South Korea

[8] Claremont Graduate University, Claremont, CA, USA

[9] Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

1 Introduction

Often, public discussions of political and social issues are plagued by deep and persistent polarization (e.g. Prior 2013; Fiorina and Abrams 2008; Großer and Palfrey 2013; Gruzd and Roy 2014; Campbell 2016; Sunstein 2007, 2017). There is obvious polarization on the national level, for example on policies about gun control and abortion (DiMaggio et al. 1996), but polarization also occurs on the small-group level, such as a faculty's division over whether students need safe spaces or a jury's polarization about whether a defendant is guilty.

Among social psychologists, the standard view is that belief polarization is the product of epistemic irrationality (Ross and Anderson 1982). The most popular view attributes polarization to biased evaluation and assimilation of evidence (see Lord et al. 1979; but also Liberman and Chaiken 1992; McHoskey 1995; Munro and Ditto 1997; Plous 1991; Sunstein 2017), while others attribute polarization primarily to motivated reasoning (Taber et al. 2009; Taber and Lodge 2006), individuals maintaining their social identity in a group (Abrams et al. 1990), or attempts to avoid uncertainty (Sherman et al. 2009; Gaffney et al. 2014). In all of these cases, the factors to which polarization is attributed are not the sorts of things that can epistemically justify or rationalize agents' beliefs, so polarization is treated as a product of epistemic irrationality.

Here we argue that there is an epistemically rational mechanism that can explain polarization. We use an agent-based model of group deliberation in which deliberation occurs by group members exchanging reasons for beliefs. By simulation, we show that polarization can be a product of agents using a coherence-based strategy for managing a limited memory. We then argue that using this coherence-based strategy is rational for limited agents of this type. Given limited memories, groups of rational agents can polarize.

Some, like Sunstein (1999), have taken it to be obvious that group polarization can be rationally produced, since it can be rational for agents to care about their reputation or group identity. What we argue is that polarization can be produced by *epistemically* rational mechanisms, a claim that Sunstein appears to deny.[1] Other models where polarization is purportedly produced by epistemically rational mechanisms have tended to emphasize the role of agents' prior beliefs and make sense only of ideal individual agents becoming more convinced of their antecedent views upon seeing ambiguous new evidence (e.g. Jern et al. 2014; Benoît and Dubra 2014; Fryer et al. 2015). In contrast, our much simpler model uses a rational mechanism and explains how groups of rational, though limited, agents can break into polarized subgroups after discussing their views.

Importantly, while our model is inspired by empirical results and uses plausible mechanisms, our aim here is only to explain how polarization could possibly be

[1] While Sunstein (1999, 2017) does think that groups can become more extreme in their beliefs via informational cascades and other mechanisms, none of those mechanisms is sufficient to break groups into polarized subgroups. Besides that, Sunstein (1999) offers no reason to think that polarization is *epistemically* rational, and his summary comments about polarization possibly producing "factual mistakes" suggests he believes it is not: "The problem [with group polarization] is … that people may be shifted, as a result of entirely rational processes, in the direction of factual … mistakes" (20).

produced by rational social belief formation. Our focus will be on *group* polarization. In social psychology, the primary focus has been on *belief* polarization, the phenomenon that occurs when two initially-disagreeing people strengthen their disagreement after seeing the same evidence (Lord et al. 1979). Sociologists and political scientists, on the other hand, have focused on *group* polarization, the formation of distinct social and political groups in societies (DiMaggio et al. 1996; Fiorina et al. 2010). In our model, group polarization is produced by the strengthening of a disagreement between members of subgroups after sharing evidence. But, because strengthening disagreement between subgroups can only occur when there's strengthening disagreement between their respective members, our results have a clear bearing on discussions of belief polarization as well.

2 Modeling group deliberation

In modeling rational deliberation, we'll start with the assumption that an agent's belief is epistemically rational when the belief stands in the right relation to the agent's epistemic reasons. This assumption sits nicely with many conceptions of reasons and rationality. Evidentialists, for example, hold that our beliefs are rational or justified when they are supported by our evidence. Evidence, on this view, is typically conceived of as reasons for belief though (McCain 2014, p. 10), so the evidentialist picture fits our starting assumption perfectly. Schroeder (2010) argues for the more general view that rational belief is belief supported by one's reasons when the reasons are construed subjectively. Schroeder (2015) makes the stronger claim that the right relation between beliefs and reasons (construed both objectively and subjectively) is a *sufficient* condition for *knowledge*. Here we assume only the weaker claim that an agent's belief is rational when it is supported by their epistemic reasons. The reader is invited to substitute their preferred term of epistemic evaluation, as long as it is one that is determined by the agent's reasons (either objectively or subjectively construed).

We'll treat group deliberation as a process whereby agents share the reasons for their beliefs. A clear example of this is the iconic jury-room scenes from *12 Angry Men* (Lumet and Rose 1957), a film in which 12 jurors must come to a joint decision about the guilt or innocence of an eighteen-year-old defendant who has been accused of murdering his father. In the jury-room, the jurors share reasons for thinking the defendant is guilty or not guilty and respond to each other's reasons. All of the jurors can hear and be heard by all of the others. In the screenplay, the jurors also get emotional, distract each other, and occasionally threaten each other, but since our aim is to model *rational* group deliberation, we don't include those elements in our model.

We employ an extremely thin conception of epistemic reasons. We model reasons as supporting belief in particular propositions with particular strengths.[2] For example, we model the fact that the store owner reported selling the odd knife used in

[2] For ease of exposition, in some places, we'll talk as though reasons support propositions or contents, rather than belief in those contents, but this is only a shorthand.

the murder to the defendant as a strong reason to believe the defendant committed the crime. The fact that the stab wound was made at an awkward downward angle and the defendant is an experienced knife fighter is modelled as a strong reason to believe the he didn't do it. This conception of reasons is very simplified. Because we model reasons as supporting a fixed content and having a fixed weight, our model doesn't naturally capture the sophisticated ways in which real epistemic reasons combine and interact. Real epistemic reasons exhibit rebutting and undercutting by other reasons, for example. But in our model, those can only be mimicked by the agent receiving a countervailing reason, one which supports belief in a contradictory content.

Despite its simplicity, this model is flexible enough to represent deliberation about a large space of propositions. In the model as we use it here, we'll assume agents are deliberating about only a single proposition, e.g. whether the defendant is guilty. We'll show that this simple case is sufficiently complex to shine light on how groups can polarize. We'll also assume that there is a fixed set of reasons relevant to each deliberation (though which reasons an agent has, and which reasons are had by anyone, will change over time).[3] Finally, we'll assume that agents are perfectly rational in that what the agent believes is determined by what is supported by their reasons. For example, if an agent has three reasons of weight 2.0 to believe P, one reason of weight 0.5 for P, and one reason of weight 3.5 for not-P, the agent will all-things-considered believe P and will do so with strength $(2.0 + 2.0 + 2.0 + 0.5 - 3.5 =) 3.0$.

There are two important dynamic aspects of the model: how agents get reasons and how they lose them. We'll assume that all agents start with the same number of reasons, though they may be different reasons. An agent's initial reasons can be thought of as representing what they initially bring to the discussion. Agents get new reasons in two different ways: either via discussion (i.e. getting them from another agent) or from the world (in *12 Angry Men*, for example, Juror 8 sees a duplicate of the boy's odd knife at a pawn shop during a break in deliberation).

Below, we consider two different kinds of group deliberation: (1) pure deliberation, and (2) deliberation with outside input. In pure deliberation, a randomly-chosen agent shares one of their (randomly-chosen) reasons with the group. All of the agents then add the shared reason (the supported proposition and its weight) to their collection of reasons, if they didn't already have it.[4] The process then repeats. In deliberation with outside input, agents receive new information from the world during the discussion. At each step of the model, each agent gets a random reason from the world (with different agents possibly getting different reasons). Then, as in pure deliberation, a randomly-chosen agent shares a randomly-chosen reason with everyone, and the process repeats. In both types of deliberation, everyone gets every reason that is shared, so there is perfect communication in the group.

Before moving on, it's worth returning to just how simple this model of reasons and deliberation is. In the model, reasons are modelled only in terms of a supported

[3] We can think of these as mirroring something fixed in the world, like the time-indexed eternal facts.

[4] We assume that the weights of reasons do not vary across agents (either because they are perfectly shared or because the weight of a reason is a priori or a matter of logic, about which our agents are omniscient). Notice that this assumption only makes our case harder to show, since if agents could reasonably assign differing weights to the same reasons, it would be easier for them to reasonably disagree.

proposition (e.g. guilty or not guilty) and a strength of support. Deliberation occurs by sharing reasons with everyone. As such, our model lacks many elements of real deliberation. It lacks expressions of emotion, expressions of desires or plans, and the personal attacks that occur in real deliberations. We consider this a virtue of our model of *rational* deliberation, but one might worry that the model cannot naturally capture some aspects of rational deliberation either. For example, our agents cannot challenge or point out mistakes in each other's reasoning. Since reasons are only for or against the proposition at hand, the agents have no way of communicating *about* the reasons or about what others say. Our model also doesn't allow agents to consider each other's reasoning or explicitly accept or reject others' claims, for example by switching someone's *modus ponens* into a *modus tollens*. Our model doesn't even allow agents to work together, for example, by one agent offering a conditional and another agent providing the antecedent.

As an idealized model, we neither intend nor expect the model to be representationally accurate or complete. Inspired by models like Schelling's (1969) model of racial segregation, our model is meant to help us understand the complex emergent phenomena of group deliberation in terms of the simple interactions of parts. Simple, idealized models allow us to understand and theorize about target phenomena in ways that would be impossible with more multifaceted models, where the effects of complex interacting elements cannot be differentiated. So while other models might fruitfully incorporate additional or more complex elements into models of group deliberation, the idea here is that a simple model is sufficient to shine light on questions of social scientific and philosophical interest.[5]

3 Rationally responding to limited capacities: ways of forgetting

As mentioned, all the agents we consider are epistemically rational in the sense that what they believe is always exactly what's supported by their epistemic reasons. Communication is also perfect in our model: every reason that is shared is heard by everyone. Despite that, we don't assume that agents have infinite memories. Of course, this is a realistic assumption, since for many topics of deliberation, no group member is in a position to know everything relevant to the topic (and in many cases, not even the entire group can know *everything* relevant). So, in some runs of the model, the agents have limited memories. Limited agents cannot remember more than their memory limit permits, except for a brief moment while they process the new incoming information that pushed them over their limit. We consider three different ways agents might manage their memory limitations.[6]

[5] Although we use these notions for quite different purposes, note the similarity of the ontology of our model to models in the hidden profile paradigm (Stasser 1988; Stasser and Birchmeier 2003; Lu et al. 2012 for a survey).

[6] Previous work has studied similar agents with limited memories. Following Hellman and Cover (1970), it's popular to model memory limitations as limitations on states of finite automata. Wilson (2014) analyses these limited automata and shows that they can polarize when the agents have differing priors. Also see Halpern and Pass (2010). These models are quite different from ours and are subject to a number of limitations discussed in Sect. 6.

The simplest way an agent might manage her limited memory is by forgetting a reason at random. For example, if this kind of agent has a memory limited to 7 reasons, when the agent gets an 8th reason, she'll lose one of the 8 reasons at random and hold on to the rest. We'll call this method of losing reasons "simple-minded." Simple-minded agents don't have the best way of handling their memory limitations, since they might, for example, forget important or conclusive reasons before inconsequential ones.

A more plausibly rational way that agents could manage their limited memories is by forgetting the reasons that are the least informative and thereby the least likely to influence their overall belief. This amounts to forgetting the reason with the lowest strength, regardless of what it is a reason for. We'll call this method "weight-minded."

Whereas weight-minded agents care only about the weight of reasons, agents might also value having coherent sets of reasons. Such "coherence-minded" agents prioritize reasons for the view that is best supported by all their reasons. When such an agent gets an 8th reason that goes over their memory limit, they prefer to forget a reason that runs contrary to the view that is supported by all 8 of their reasons. Coherence-minded agents (like their weight-minded counterparts) also value basing their views on the most informative reasons they have. This means that when they face memory limitations, they forget reasons for opposing views starting with the least strong. Only after that do they forget reasons supporting their own view (again, starting with the least strong).

Later we'll argue that coherence-mindedness is an epistemically rational way to manage one's memory limitations. First though, we'll show why coherence-mindedness is interesting. The dynamics of deliberation among groups of coherence-minded agents turn out to be dramatically different from the dynamics of groups that use the other strategies.

4 Model simulation results

We simulated our model in Netlogo using groups of 50 agents, in a world with 500 total reasons (only some of which are held by agents), and assuming limited agents can retain only 7 reasons at a time.[7] Reasons were randomly assigned to support one of two propositions (P or not-P, e.g. *the boy is guilty* or *the boy is not guilt*y). The strengths of the reasons were randomly assigned by an exponential distribution with mean 1. Because of the way reasons were created we should expect the vast majority of reasons to have a low weight (i.e. they're small points in favor of a particular view, like the boy having the hair color reportedly seen by a witness). The qualitative results described here are robust against changes of this distribution, as long as they save that general characteristic.[8]

[7] The reader should think of these as the agents' reasons that bear on the relevant proposition, not all of the reasons they have. We use 7 as the limit following Miller (1956), though recently Cowan (2001) has argued that the number should be 4. See the discussion of the robustness of this result below.

[8] Our results are robust for various other distributions of reasons that have a similar qualitative characteristic. We don't discuss distributions that make put more reasons strongly on either side, since polarization would be less surprising in those cases.

To measure how polarized the simulated groups are before, during, and after deliberation, we use four measures inspired by Bramson et al. (2016, 2017). The simplest measure, which actually measures a lack of polarization, is *time to convergence*. The time to convergence is the number of steps of the model it takes for the population to converge for the first time on a single view (i.e. everyone having the same belief content) or on a particular collection of reasons (i.e. everyone having the same belief content with the same reasons). If groups converge, either to a view or a set of reasons, then they are not polarized. Of course, not every group will converge, but when they do, lower times to convergence will indicate less room for polarization.

The second measure we use is *subgroup divergence*. In our model, there will often be two subgroups: those who think the boy is guilty and those who don't. Subgroup divergence is a measure of how far apart the two subgroups are, which we measure in terms of the distance between the means of the strengths of the agents' beliefs in each subgroup. Intuitively, when the subgroup divergence is high, the two subgroups strongly disagree, and the population is more polarized.

Third, we look at *subgroup consensus*. Whereas subgroup divergence tells us about how the two groups of agents relate to each other, subgroup consensus tells us something about what's going on inside the subgroups, namely how tight-knit each subgroup is. This is measured in terms of variation from the subgroup mean. Intuitively, two distinct groups with low subgroup consensus are not as polarized as two similar subgroups with higher internal consensus. So if a population has a high subgroup divergence (the second measure) and a high subgroup consensus (the third measure), it is extremely polarized.

The fourth measure tracks the relative size of the polarized subgroups. Intuitively, a community is more polarized when its subgroups have close to equal size. Bramson et al. (2016) call this 'size parity,' but the complexity of Bramson et al.'s formal measure of size parity is not needed here. Instead we'll give summary statistics about the proportions of runs in which different relative sizes of subgroups appear.

4.1 Results from pure deliberation

We start with the simplest case: agents with unlimited memories in pure deliberation (with no outside input). We would expect this population to ultimately converge on a single view (the one warranted by the collection of all reasons held by any agent initially). And that is what we see in the simulation runs: all 50 of the agents eventually end up with the same set of reasons, giving them the same view with the same strength. This convergence often takes an extremely long time though (on average, 61,291 steps to a shared set of reasons).[9] Even before converging, these agents do not display much of what intuitively looks like polarization. We can see this by looking at the time slices of a histogram of beliefs and strengths in Fig. 1. In

[9] Of the 1000 runs done to test this, 32 of the runs didn't converge in fewer than 100,000 steps, the limit we set in testing. These runs would have converged, given more time. So the real averages are even higher.

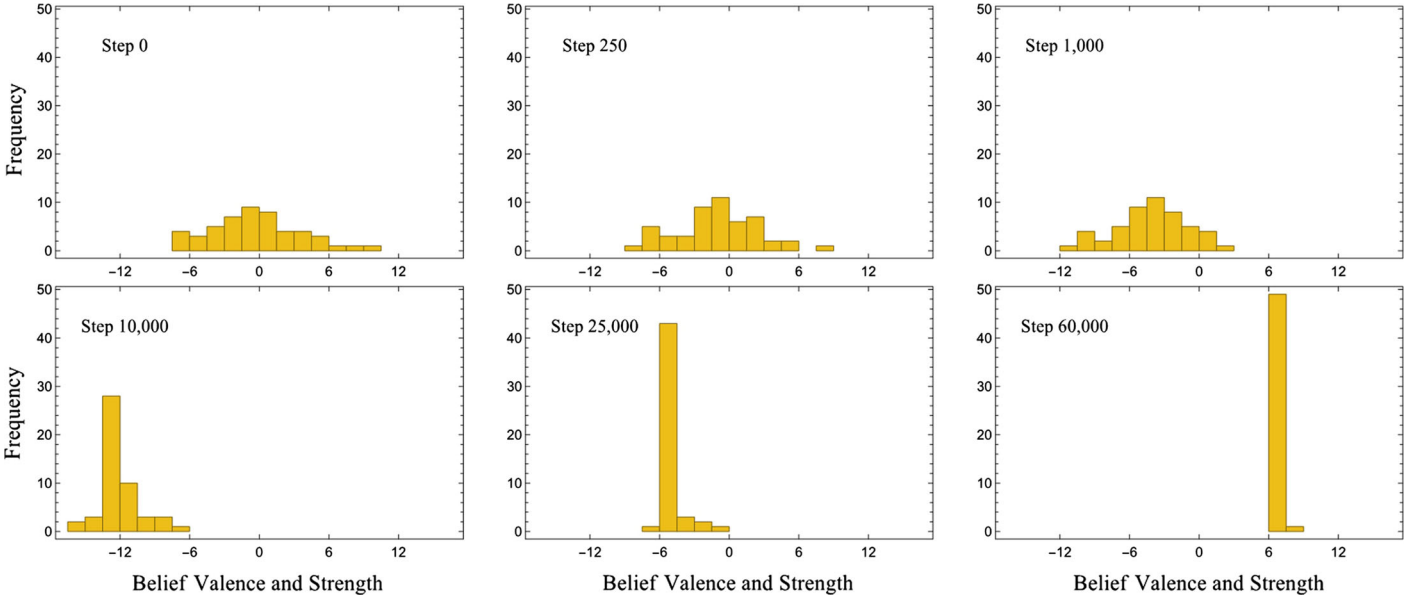

Fig. 1 Histogram of beliefs and strengths for a typical run with unlimited agents

the histograms, the strengths of agents' beliefs for those who believe P is plotted on the positive side, and the strengths of the agents' beliefs for those who believe not-P is plotted on the negative side.

In pure deliberation, collections of simple-minded agents who can only remember 7 reasons at a time act qualitatively like collections of unlimited agents. They converge on a single view and strength for that view, but they do so much more quickly (on average, 3633 steps to a shared set of reasons). What explains the quicker time to convergence? Groups of simple-minded agents start out with a lot of different reasons spread through the population. With many reasons and a relatively small memory limit, this means very few reasons are had by multiple people at the beginning. After a reason is shared though, that particular reason is likely to be remembered by many agents, since every agent gets the shared reason before they forget a random reason. So each time a reason is shared, it can be expected to become more widely held, and the total diversity of reasons held decreases (with the shared reasons being more prevalent than the unshared ones). This makes the group converge on a set of reasons more quickly than unlimited agents.

Groups of weight-minded agents, who forget the weakest reason regardless of what it's a reason for, also converge to sharing all their reasons in pure deliberation. In this case, it's because the lowest-weight reasons are systematically forgotten by all agents, so all of the highest-weight reasons end up being heard (and remembered) by everyone. That mechanism results in even quicker convergence than groups of the other two kinds of agents (on average, 794 steps to a shared set of reasons).

Groups of coherence-minded agents, unlike the other kinds, in general do not converge on either a shared set of reasons or an overall view in pure deliberation. In 1000 runs, only 36 runs converged on a set of reasons and only 45 converged on a view.[10] The dynamics and stable end-states of these groups are also dramatically different from the other kinds of groups. Groups of coherence-minded agents typically break into two smaller subgroups, one on each side of the issue. Those subgroups separate from each other by becoming more entrenched in their view. Finally, the subgroups converge internally by coming to share the same reasons. The two subgroups then settle on a set of reasons for their respective views, and the community becomes stably polarized. This dynamic can be seen in the series of histograms representing a typical run of the model in Fig. 2.

The dynamics for coherence-minded agents can be better understood in comparison to the others by comparing the subgroup divergence and consensus measures. Figure 3 shows the change in subgroup divergence over time for 100 runs of each of the different kinds of groups. When subgroup divergence goes down, the subgroups are getting closer together, so higher subgroup divergence intuitively indicates higher polarization.

[10] These numbers assessed in different sets of 1000 runs. We stopped the run if convergence didn't happen by 100,000 steps, since most converging runs did so very quickly (fewer than 350 steps for convergence on a view and fewer than 1500 steps for convergence on a set of reasons).

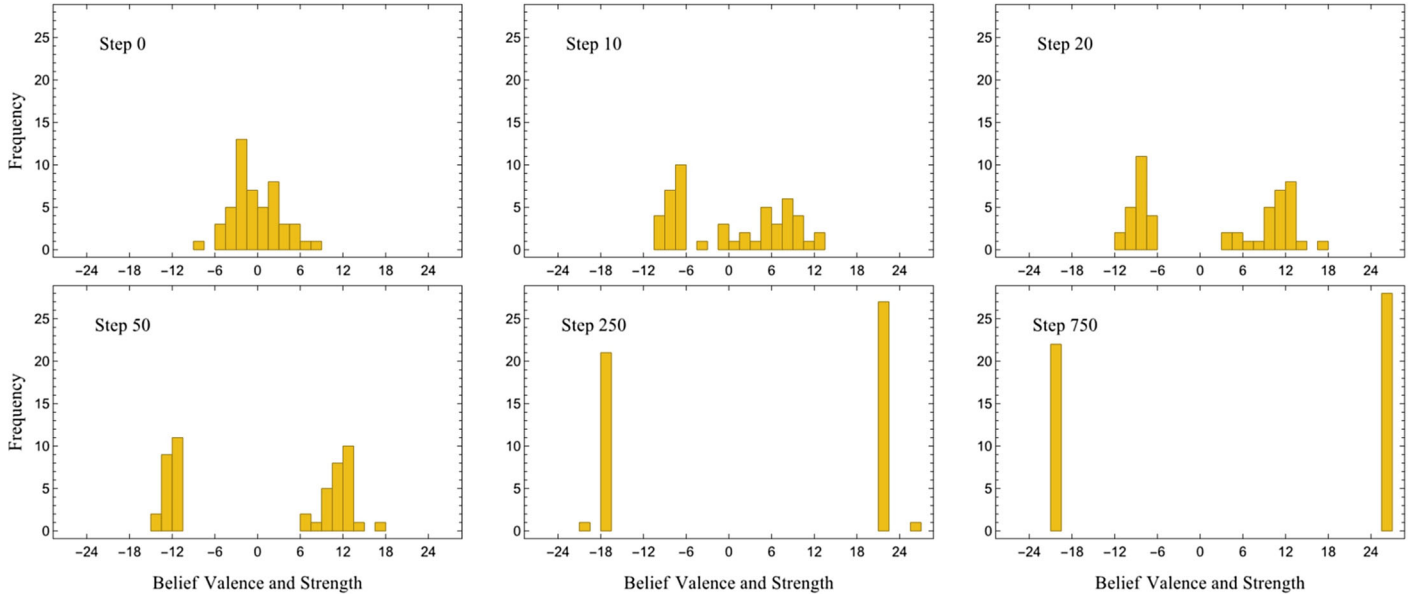

Fig. 2 Histogram of beliefs and strengths for a typical run with coherence-minded agents

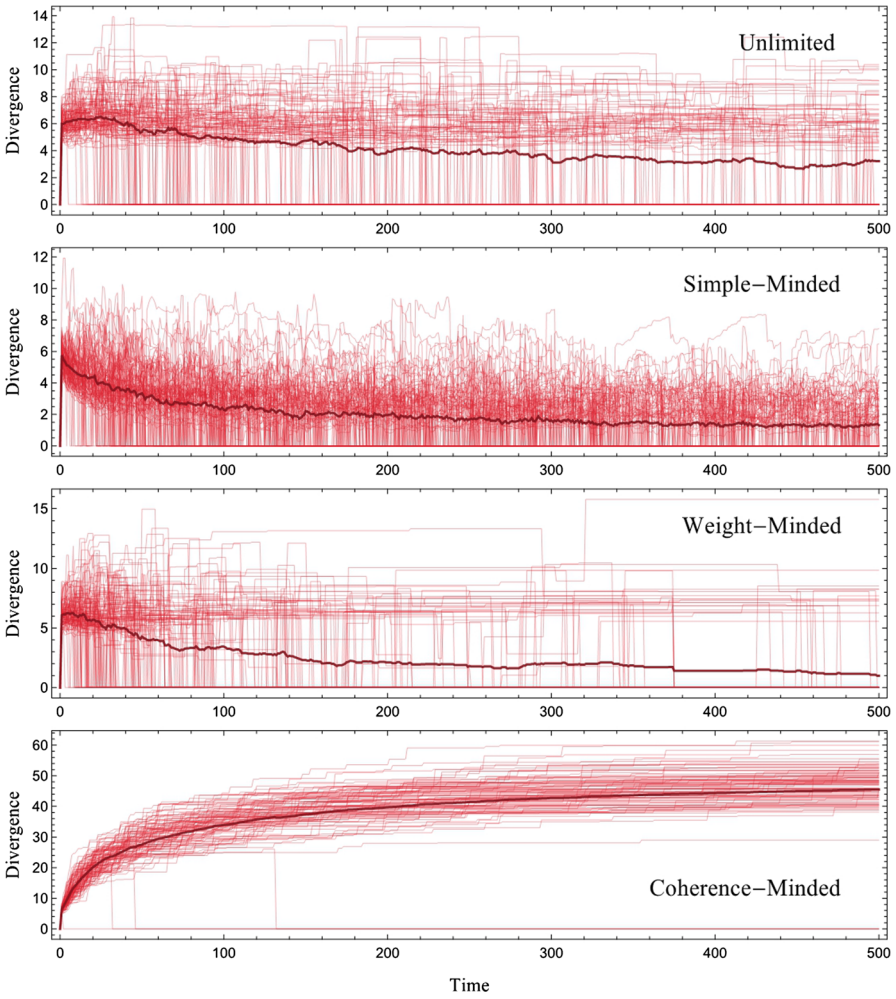

Fig. 3 Subgroup divergence per run per step for 500 steps of 100 runs. Note the differing scales of the y-axis. Subgroup divergence is zero only when there is one group, and subgroup divergence increases as the groups separate. The darker line is the average at each time step of all 100 runs

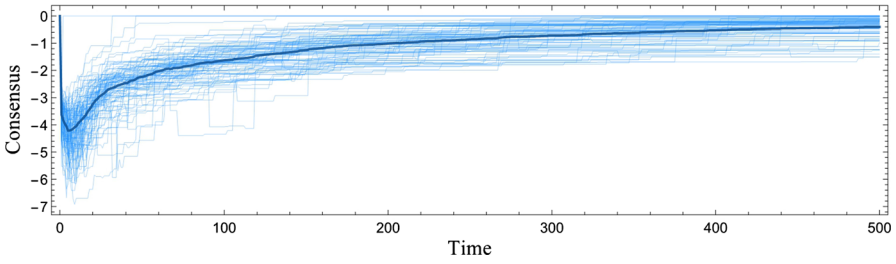

Fig. 4 Subgroup consensus per run per step for 500 steps of 100 runs of coherence-minded groups in pure deliberation. Zero is maximal in-subgroup cohesion

As the graphs show, subgroups of coherence-minded agents tend to separate over time, unlike subgroups of the other kinds of groups who pull together (depolarize) over time.

Subgroup consensus over time for groups of coherence-minded agents appears in Fig. 4. What Fig. 4 shows is that the two subgroups of coherence-minded agents become more internally cohesive over time. Combining this with the information from Fig. 3 shows that over time, these agents disagree with the other subgroup's members more strongly and agree within their own subgroup more strongly, which indicates high degrees of polarization.

When we look at the sizes of the subgroups, we see that coherence-minded agents tend to polarize into groups of roughly equal size: Over 71% of the time when polarization occurs, at least 20% of the population is in the minority group. Over 30% of the time, the minority group is 40% of the population, and 20% of the time, the minority group is three or fewer agents shy of being half the population. It is in less than 2.1% of the runs that the minority is only one or two agents. In cases of pure deliberation, therefore, we see consistent and significant polarization of coherence-minded agents, which contrasts starkly with the dynamics of groups of the other kinds of agents.

4.2 Results from deliberation with outside input

In contrast to pure deliberation, for deliberation with outside input, each agent gets a (possibly different) reason from the world at each step of the model before a randomly-selected agent shares a randomly-selected reason with everyone.

Here again, agents with unlimited memory always converge. In this case, they do so by gathering all the reasons, which they can do quickly (on average, 372 steps, compared to 61,291 in pure deliberation). With cognitively limited agents, the injection of reasons from the world doesn't have the same impact, since limited agents can't retain the complete set of reasons like unlimited agents do. In the case of simple-minded agents, because the agents are constantly getting new reasons from the world and because they forget those reasons randomly, groups rarely end up agreeing for long and they almost never share a set of reasons. Though the agents never converge, the population also isn't polarized. Instead, the agents simply meander around the space of beliefs, with no distinctive subgroups forming (Fig. 5).

Measures of subgroup consensus for these runs also indicate that these agents do not solidify around their view over time. As such, though groups of simple-minded agents fail to converge with outside input, they can't be seen as polarized either.

As with groups of unlimited agents, adding outside input to the deliberation of weight-minded agents significantly decreases the time to convergence on a set of reasons (on average, 140 steps with outside input, compared to 794 in pure deliberation). The quicker convergence here can be explained by the outside input providing access to the reasons more quickly. Since weight-minded agents all share a preference about which reasons to forget, that they can access the reasons more quickly means that they will converge more quickly on the same shared set of reasons.

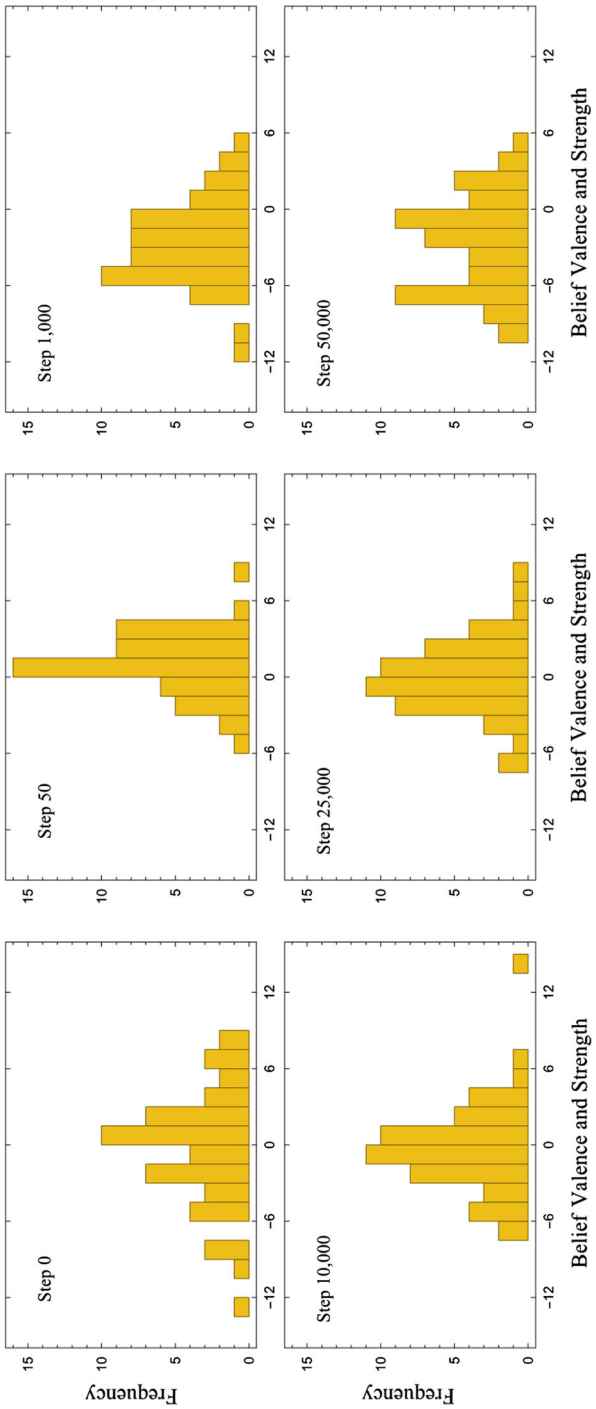

Fig. 5 Histogram of beliefs and strengths for a typical run with simple-minded agents with outside input

For coherence-minded agents, the story is different. The addition of outside input increases their degree of polarization. In pure deliberation, groups of coherence-minded agents sometimes converged on a set of reasons or a view (36 and 45 runs, respectively, out of 1000). In cases of deliberation with outside input, none of the 1000 runs converged at all.[11] In every run, two polarized subgroups emerge. With outside input, the subgroups of the polarized population reach an internal consensus more quickly too, as confirmed by the subgroup consensus data. After 200 steps, for example, runs with outside input have a subgroup consensus of roughly 0.5 on average, whereas in pure deliberation, it was 1.0. (Lower numbers indicate more cohesive subgroups.) Also, subgroup divergence is higher with outside input (from roughly 35 on average in pure deliberation after 200 steps to roughly 60 with outside input). This means that the subgroups more strongly disagree when they have outside input. Finally, the size of the polarized subgroups tends to be more equal with outside input than in pure deliberation. With outside input, 20% or more of the population is in the minority in over 92% of the runs (compared to 71% of runs in pure deliberation). 40% or more of the population is in the minority in 44.7% of runs (compared to 30% of runs in pure deliberation). And 29.9% of runs leave the majority within 3 agents of being half of the population (compared to only 20% of runs in pure deliberation runs). These numerical results are summarized in Table 1.

4.3 Results, robustness, and reality

What the above results show is that, in contrast to groups of the other kinds of agents, when outside input is added to deliberating groups of coherence-minded agents, the groups become more polarized. For unlimited and weight-minded agents, the addition of outside input increases the speed at which the groups converge, which intuitively points towards less polarization. Simple-minded groups don't exhibit this effect, but the outside input doesn't seem to push them towards polarization either. These qualitative results proved resilient in all of our robustness tests, which included combinations of changes in the number of agents involved (from 3 to 200), the total number of reasons (from 100 to 1000), the number of reasons agents can remember (from 3 to 25), and whether that number is homogenous or heterogeneous in the population. So, all told, when we add outside input to the deliberation, we see an increase in polarization for coherence-minded agents while, in most other cases, we see a clear decrease in polarization. This indicates that the polarization is a result of using a coherence-based mechanism of memory management.

We know that groups of agents with unlimited memories will always converge in this model, so it's essential to our result that the agents are limited. In the results described above, individual agents could only remember 7 of the 500 total reasons. So it's natural to worry that if the result only obtains for very limited

[11] We ran 1,000,000 runs to see if consensus is ever reached in this setup. Of those, less than 0.01% of them converged on a view and less than 0.005% converged on a set of reasons. When those cases did converge, it always happened in the first 9 steps of the model (many in the first or second step), which indicates a rare combination of initial conditions and early steps is required.

Table 1 Comparison of the mean time to convergence on a set of reasons, size parity (in scenarios in which polarization does occur), subgroup divergence, and subgroup consensus for all agent types in both pure deliberation and deliberation with outside input

Agent type	Pure deliberation				Outside input			
	Convergence	Size parity	Divergence	Consensus	Convergence	Size parity	Divergence	Consensus
Unlimited	61,291 steps (SD 15,063)	N/A	Low and decreases	N/A	371 steps (SD 5.6)	N/A	Low and decreases	N/A
Simple-minded	3633 steps (SD 3096)	N/A	Low and decreases	N/A	Never	N/A	Low and stable	N/A
Weight-minded	794 steps (SD 405)	N/A	Low and decreases	N/A	140 steps (SD 38)	N/A	Low and decreases	N/A
Coherence-minded	Almost never	Usually near even split	High and increases	High Increases	Never	Almost always near even split	Very high and increases	Very high and increases

agents, our result couldn't shine light on any real social groups. The result is quite general though. As mentioned above, we observed all of the same qualitative features of every variation of the model when there were only 100 reasons and agents could remember 25. Even tests where agents can remember half of 100 reasons show a significant amount of polarization: In 1000 runs of each, 68.2% of cases of pure deliberation polarized, as did 81.5% of cases of deliberation with outside input.[12]

More generally, we might ask how realistic we can expect this model to be. Of course, the model is highly idealized and not intended to faithfully represent real human deliberation. That said, the mechanism of action in this model is, on the very broad level, quite similar to the mechanism of action supported on empirical grounds by Lord et al. (1979) and the large literature that has followed. At base, what's happening with agents in our model and that literature is that the prior evidence that agents have is affecting how they react to their later evidence. In both cases, agents generally end up retaining information that favors what they antecedently believed at a higher rate than information that shows against their view. The particular mechanisms the agents use to do that differ though, and we think that difference is epistemically important. Whereas our agents are being rational in light of their limitations, Lord et al.'s aren't. This is what we argue for in the next section.

5 Coherence-mindedness is epistemically rational for limited agents

In contemporary epistemology, discussions of rationality are usually about what beliefs or credences are rational to have in a given situation. Methods of managing memory are not usually thought of as the type of thing that can be epistemically rational or not, perhaps because it's generally thought that we lack control over what we forget. We consider this line of reasoning to be mistaken. As is well-argued in the literature, we often cannot control what we believe, but that doesn't undermine construing beliefs as rational.[13] Moreover, in many respects, we seem to control our memories at least as well as we control our beliefs. If we want to remember some fact, for example, we can repeat it to ourselves often or write it down. If we want to remember (or forget) certain *kinds* of facts, we can train ourselves to pay more (or less) attention to those kinds of facts in our daily lives. So, although we don't directly control what we forget, that doesn't undermine the idea that methods of managing memory can be more or less rational.

So what factors bear on the rationality of memory management? If we thought simple-mindedness were rational, then we'd be lacking for an epistemic critique of a juror who voluntarily remembered barely-relevant testimony about an alleged assailant's upbringing over eyewitness testimony about the assault. So surely, when

[12] In fact, it would be possible for groups to polarize even if their memory were limited to only 1 fewer than the total number of reasons, but we'd expect this to occur quite infrequently.

[13] We point the reader to Vitz (n.d.) for background on the connection between doxastic volunteerism and evaluation.

we're confronted with memory limitations, rationality requires us to put some priority on remembering more informative over less informative evidence. This suggests that weight-mindedness might be the most rational way to manage a limited memory, since weight-mindedness tells agents to prioritize remembering their strongest reasons.

Weight-mindedness has an odd consequence though. In some cases, a weight-minded agent must see her future self as both having a less accurate belief and having it on the basis of an inferior set of evidence. To see why, consider this example: Suppose agent Ada has a memory limit of 4 reasons and currently has a full memory containing reasons of strength 6 and 7 for P and reasons of strength 6 and 8 for not-P (so that her overall belief is not-P with strength 1). If Ada then receives a new reason of strength 5 for P, her evidence will overall support believing P with strength 4. Once she applies the weight-minded forgetting rule, she'll forget that new reason (since it's the weakest), and she'll end up believing not-P again. What's the problem here? The problem is that $\text{Ada}_{\text{current}}$ is in a position to evaluate whether P on the basis of all 5 pieces of evidence, which is all of the evidence $\text{Ada}_{\text{future}}$ has plus more. So from $\text{Ada}_{\text{current}}$'s perspective, $\text{Ada}_{\text{future}}$ will not only have the wrong belief about P, but will also be in a worse position to make that judgement. If she's weight-minded, therefore, Ada will see her future-self as both wrong about the question and ignorant of the relevant evidence. Given that there's an alternative, rationality surely can't require us to act like that.[14] And there is an alternative. Coherence-mindedness never requires us to manage our memory that way.

It's natural to think that coherence-minded agents are biased in the same way as the agents described by Lord et al. (1979) and their followers (e.g. Liberman and Chaiken 1992; McHoskey 1995; Munro and Ditto 1997; Plous 1991). According to that literature, real agents biasedly evaluate and selectively adopt evidence that supports their antecedently held view and selectively reject evidence contrary to that view. Might coherence-minded agents be similarly biased? No. What makes Lord et al.'s agents biased is that how they treat new reasons is a function of what they *antecedently* believed. Coherence-minded agents forget a reason as a function of what is supported by all of their reasons, including both the ones they had before and the new ones they've received that pushed them over their memory limit. It's difficult for Lord et al.'s agents to change their minds when they hear a new reason, because they'd only incorporate the reason if it agrees with them. Our agents are comparatively open-minded. Unlike Lord et al.'s biased agents, our agents never misjudge the content or strength of their evidence, nor do they misprocess evidence they receive. Our agents incorporate new reasons before deciding what to forget, and as such they aren't irrationally stubborn like biased assimilators.

A number of rationality considerations favor coherence-mindedness. Consider coherence-mindedness in terms of a picture of epistemic normativity in which coherence of doxastic states has value. Of course, among epistemic internalists, it is

[14] If the reader hasn't already given up on the rationality of simple-mindedness, she is encouraged to notice that simple-mindedness is subject to this same kind of worry.

often assumed that coherence of one's beliefs is a necessary condition of their justification or rationality (e.g. Bonjour 1980; Lehrer 1990), which naturally gives support to being coherence-minded when limited. But even many authors who reject internalism, like Sosa (1985) and Foley (1993), uphold coherence as an epistemic value. Similar ideas appear in discussions of Bayesianism, which, following Ramsey (1926), defend the idea that coherence is rationally required for credences, (e.g. Joyce 1998).[15] If we assume that having coherent doxastic states has epistemic value, coherence-mindedness is a natural way of creating that value. Forgetting a reason for an opposing view will always leave us with greater support for our view than the alternative. Given our limitations, we're bound to forget reasons. Given that we must forget a reason, coherence considerations push us toward keeping ones that favor the view supported by our evidence over all.

It's important to notice that coherence-based considerations themselves are insufficient to rationalize the coherence-mindedness forgetting rule. If coherence were the only value, agents would be required to forget the *strongest* reasons for opposing views (not the weakest as coherence-mindedness dictates), since doing so would make one's doxastic state even more intuitively coherent. Agents who were *only* concerned with coherence would be overly biased towards their own views and could be expected to thereby intentionally mishandle their evidence, for example, by forgetting *all* of the opposing evidence. Being rational also requires one to respect one's evidence by basing one's beliefs on one's most informative reasons. When an agent has many considerations in favor of an opposing view and she has a limited memory, rationality also pushes her to forget only the least informative ones. With coherence as one epistemic virtue among many, coherence-mindedness in the form outlined looks like a rational strategy for managing limited memory.[16]

Another reason to think coherence-mindedness is rational is that it's suggested by a plausible story about what it is to take evidence to be misleading. Let's start with an example. Suppose you rationally believe that Bob did not commit a murder even in light of the fact that Bob's fingerprints were found on the murder weapon. Also suppose you believe that finding someone's fingerprints on a murder weapon is a good way to find out that they committed the murder. If you maintain your rational belief that Bob did not commit a murder, your belief that his fingerprints were on the weapon, and your belief about the evidential import of fingerprints, rationality requires that you take the fingerprint evidence to be misleading. This is an instance of a general principle that if you rationally believe P, rationality requires you to either treat evidence that not-P as misleading or stop believing P. Let's call this claim *rational dogmatism.*

Rational dogmatism supports thinking that coherence-mindedness is rational for limited agents. To see why, notice that judging a piece of evidence *e* to be

[15] See, for example, Murphy (2016) for a discussion of how foundationalists often appeal to notions of coherence. Cohen (1984, pp. 283–284) also argues that "justification" and "rationality" are synonymous as used by epistemologists. As such, coherentists' claims about justification ought to translate to rationality as well.

[16] That said, groups of agents following that more extreme rule still polarize.

misleading requires thinking that *e* supports a proposition *P* and thinking that one shouldn't believe *P* on the basis of *e*. So, if we have to forget a reason due to a limited memory, rationality should push us to forget the apparently misleading ones before the others, since we judge those reasons to be defective, i.e. not reasons we should base our beliefs on. (Notice the similarity of the reasoning here with the intuitive reasoning behind thinking that weight-mindedness was rational: in both cases, it seems like we should give up the reasons that it would be less good to base one's beliefs on.) If we again add in that, other things being equal, it is rational to keep more informative reasons when given a choice, it follows that when faced with a memory limitation, rational agents will forget the weakest reasons that oppose their view.[17]

Although the argument here is reminiscent of the Kripke–Harman Dogmatism paradox (Harman 1973, p. 148), it is subtly different in ways that don't lead to the same paradoxical outcome. The Kripke-Harman paradox starts with the idea that if one knows P, then one must regard any future evidence that goes against P as misleading and thereby disregard that evidence. Following this line leads to the conclusion that knowing P licenses one to conclude that *any* evidence that goes against P is ignorable. Doing this doesn't make an agent rational; it makes him arrogant. Our reasoning does not lead to the same outcome. The Kripke-Harman paradox arises only when agents ignore any evidence that conflicts with what they know (or rationally believe, in our case). But our agents don't ignore views that oppose their own. Our agents first determine what is supported by all of their reasons, including reasons that support both sides. Only in light of that entire set of reasons do they forget any reasons when forced to by their memory limitation. With our agents, new reasons that conflict with their antecedent view might in fact convince them to change their view. The impossibility of that change is what makes the Kripke-Harman agents irrational.

Before closing, consider one more worry one might have about coherence-mindedness. Whereas weight-mindedness required Ada to radically change beliefs when forgetting, leaving their future self in an avoidably worse-off position, coherence-mindedness appears to have the opposite problem. Consider Cada, who can remember 7 reasons and currently has reasons of weight 10, 2, 1, 0.5, and 0.5 in favor of P and reasons of strength 7 and 6 against, which leaves Cada believing P with strength 1. Suppose Cada receives a reason of strength .5 in favor of P. Coherence-mindedness would require Cada to forget the reason of strength 6 against P in order to remember this new very weak reason for P. But, isn't that the wrong result? Isn't Cada placing too much value on coherence in forgetting the strong reason for not-P just to remember a weak reason for what they weakly believe?[18]

[17] One might think that if a reason's strength is a measure of how misleading it is, agents should forget the strongest opposing reasons (not the weakest, as coherence-mindedness requires). But, strength is a measure of how much the reason supports a view, and whether it is misleading is a question of whether it supports the truth. If one thinks that rationality requires that we drop the strongest opposing reason, that rule would still produce polarization, and so our primary conclusion still holds.

[18] We're thankful to an anonymous reviewer for bringing this case to our attention.

As we mentioned above, we think of coherence as just one epistemic virtue among many. In that vein, there are two ways of thinking of what went wrong in Cada's case. First, you might think that Cada shouldn't have placed any weight on coherence given how weakly their overall evidence showed in favor of their belief before they got the new evidence. A proponent of this story would hold that using coherence-minded memory management is only warranted when one's belief is sufficiently strong to begin with, and otherwise, when one's evidence is more mixed, one should be weight-minded. Alternatively, one might think that what went wrong in Cada's case is that Cada shouldn't have put coherence lexicographically prior to other rational considerations. Instead Cada should have treated the reason's coherence with the others as a factor in keeping it, one that might be outweighed by its relative strength. Both of these alternative forgetting rules countenance coherence as a virtue but allow that there can be cases where we should forget weak reasons on our side to save stronger ones for alternative views.

Notice though that any memory-management rule that gives some weight to coherence in deciding what to remember should admit of some cases of polarization. This is because, for any such rule, there must be cases where two agents with different beliefs should treat their evidence differently, and in virtue of that, end up moving in opposite directions. For the rules just described, this is easy to test. We implemented the two rules mentioned above in the model. In the first case, our agents acted weight-mindedly unless they had quite strong reason overall for their belief, in which case, they switched to being coherence-minded.[19] In 1000 runs of pure deliberation, 68.4% of those resulted in polarized groups. In 1000 runs of deliberation with outside input, 87.7% polarized. In the second case, our agents didn't universally prefer to preserve reasons in favor of their view. Instead, they preserved reasons for their view when they weren't much less weighty than the weakest reasons for the alternative view.[20] Here we found that polarization still occurred in 49.3% and 84.1% of 1000 runs of each of pure deliberation and deliberation with outside input, respectively.

So while coherence-mindedness can be given a defense by treating coherence as an epistemic value, it's not essential to our argument that coherence-mindedness is the unique best way to account for the rational value of coherence. Any rule that accounts for the rational value of coherence will admit of some cases of rational polarization.

[19] In our actual tests, we assumed that an agent had a strong enough belief to be coherence-minded when the strength of their belief was at least a quarter of the weighted strengths of all of the reasons.

[20] In our tests, we implemented this by asking agents to treat reasons in favor of their view as 2 times as important as reasons against and then asked the agents to forget the weakest reason in that new ranking. So a reason of weight 1 in favor of their view would be saved over a reason of weight 1.5 against, but not against a reason of weight 2.5 against.

6 Other accounts of "rational" polarization

There are many models of group polarization in the social science literature (e.g. the class of models inspired by Axelrod (1997), and the models from Hegselmann and Krause 2002, 2005, 2006). Bramson, et al. (2017) shows that there are problems with using those models to understand many observed forms of polarization. We'll rely on that critique of those models as descriptive models of polarization, and here we'll focus on the contrasts between ours and other models of purportedly *rational* polarization.

In one recent class of Bayesian models, rational polarization is purportedly derived from agents having different prior beliefs, which causes them to polarize after seeing the same evidence. Jern et al. (2014) and Benoît and Dubra (2014) give models of this sort. Fryer et al. (2015) gives a more realistic model in which agents biasedly interpret ambiguous evidence and then incorporate what was interpreted (c.f. Lord et al. 1979).

We take it that Fryer et al.'s agents are irrational in the same way that Lord et al.'s are, but more generally, we worry that polarization that's only due to agents having different priors may not necessarily be as rational as these authors assume. One might worry, for example, that having the needed priors itself might be irrational, if for example one of them is counter-inductivist. As Talbott (2016) points out, it is only the most extreme subjective Bayesians (a small group of theorists) who say that *any* prior counts as rational. Moreover, any group polarization (as opposed to belief polarization) that appears in these models won't be explained by the agents interacting in any way—it's simply sets of individuals who happened to have similar priors. A further story about why groups of people would share the same priors would be needed to explain group formation. Our model points towards a more nuanced picture in which group interaction is crucial to group polarization.

Halpern and Pass's (2010) model presents a different picture of how agents might polarize. In their model, limited agents approximate probabilistic inference with costly computation. Using expected utility theory and treating computation as costly, agents choose a way to compute predictions in light of incoming information. Halpern and Pass show that when these agents try to optimize the trade-off between expected accuracy and computational costs, they can polarize.

One similarity between our model and theirs is that group polarization is treated as phenomenon of non-ideal (though fully rational) agents. In other respects, the models are quite different. In Halpern and Pass's model, computation is costly but possible, whereas in ours, agents face no computation costs, only fixed memory limits. Also, in their model, group polarization only occurs by independent, non-interacting individuals having the same beliefs. Finally, Halpern and Pass's agents use expected (epistemic) utility theory to form their beliefs. But, as Greaves (2013) shows, ideal agents using expected epistemic utility theory aren't plausibly rational. If that result carries over to non-ideal agents, we shouldn't expect Halpern and Pass's agents to be epistemically rational either.

Finally, in the philosophy literature, Kelly (2008) argues that belief polarization can occur when agents' different histories have different causal effects on their evidence. In both our formal model and Kelly's qualitative picture, agents' evidence exhibits a kind of path-dependence, and in that sense, our argument is quite complementary to Kelly's. The upshots of our work and Kelly's are quite different though: whereas Kelly describes how rational, sophisticated, ideal agents might polarize, we show that rational, simple, and limited agents also polarize. Unlike Kelly, who focuses on individual believers, our model shows how the dynamics of many individual agents can generate subgroups of the population who agree with each other but disagree with members of other subgroups.

Altogether then, our model offers a distinctive explanation of how simple and epistemically rational agents with limited memories can interact to form polarized groups.

7 Further implications and conclusion

What we argued above is that, for memory-limited agents, being coherence-minded is a rational memory-management strategy, and groups of coherence-minded agents can be expected to polarize into subgroups that both steadfastly disagree and become more internally cohesive. The possibility of fully rational agents polarizing has implications for many areas of social and political philosophy, political science, sociology, and public policy. Our results support Sunstein's (2002) claim that real polarization threatens the central idea of the deliberative democracy literature (Gutmann and Thompson 1996; Landemore 2013; Knight and Johnson 2011), and contest Landemore's (2013, p. 138) and Knight and Johnson's (2011, pp. 124–125) objection that polarization only occurs in quite unideal or uncommunicative societies. In philosophy of science, our model and results can also shine light on model-based discussions of the division of cognitive labor in science (e.g. Kitcher 1990; Strevens 2003; Hegselmann and Krause 2006; Zollman 2007; 2010; Grim and Singer et al. 2013). Our model of agents sharing reasons in deliberation is more sophisticated than extant models of scientific discussion while still being theoretically parsimonious and tractable. Our results also suggest that current views of when scientific consensus is rational need reconsideration. In contrast to the popular views, our results show that rational scientists can disagree after sharing evidence, even without extreme priors (as they would need in Zollman's models; but see Bruner and Holman forthcoming).

Finally, notice that this discussion also undermines a natural line of thought about group polarization in today's societies. In real disagreements, it's common for parties on both sides to see the other side as blind to truth, epistemically corrupt, or simply irrational. The motivation might be that if one group is responding to their reasons correctly and sharing them with others, then if the others still disagree, the others must be in the wrong. We show that this line of thought is mistaken. Limited

agents, like real humans, can be epistemically rational and still persistently disagree in ways that produce political and social polarization.[21]

References

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology, 29*(2), 97–119.

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution, 41*(2), 203–226.

Benoît, J. P., & Dubra, J. (2014). *A theory of rational attitude polarization*. Available at SSRN 2529494.

Bonjour, Lawrence. (1980). Externalist theories of empirical knowledge. *Midwest Studies in Philosophy, 5,* 53–74.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Fisher, S., Sack, G., et al. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science, 84,* 115–159.

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., et al. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology, 40*(2), 80–111.

Bruner, J. & Holman, B. (forthcoming). Complicating consensus. In Garbayo, L. (Ed.), *Expert disagreement and measurement: Philosophical disunity in logic, epistemology and philosophy of science*. Dordrecht: Springer.

Campbell, J. E. (2016). *Polarized: Making sense of a divided America*. Princeton: Princeton University Press.

Cherniak, C. (1981). Minimal rationality. *Mind, 90*(358), 161–183.

Cohen, S. (1984). Justification and truth. *Philosophical Studies, 46*(3), 279–295.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have Americans' social attitudes become more polarized? *American Journal of Sociology, 102,* 690–755. https://doi.org/10.1086/230995.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton: Princeton University Press.

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science, 11,* 563–588.

Fiorina, M. P., Abrams, S. J., & Pope, J. (2010). *Culture war?*. New York, NY: Pearson Longman.

Foley, Richard. (1993). *Working without a net: A study of egocentric epistemology*. Oxford: Oxford University Press.

Fryer, R. G., Jr., Harms, P., & Jackson, M. O. (2015). *Updating beliefs when evidence is open to interpretation: Implications for bias and polarization*. Working Paper. Retrieved from http://scholar.harvard.edu/fryer/publications/updating-beliefs-ambiguous-evidence-implications-polarization.

Gaffney, A. M., Rast, D. E., III, Hackett, J. D., & Hogg, M. A. (2014). Further to the right: Uncertainty, political polarization and the American "Tea Party" movement. *Social Influence, 9,* 272–288.

Greaves, Hilary. (2013). Epistemic decision theory. *Mind, 122*(488), 915–952.

Grim, P., Singer, D. J., Fisher, S., Bramson, A., Berger, W. J., Reade, C., et al. (2013). Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme, 10*(4), 441–464.

Großer, J., & Palfrey, T. R. (2013). Candidate entry and political polarization: An antimedian voter theorem. *American Journal of Political Science, 58*(1), 127–143.

Gruzd, A., & Roy, J. (2014). Investigating political polarization on Twitter: A Canadian perspective. *Policy and Internet, 6*, 28–45.

[21] We take ourselves to be adding to the literature that emphasizes the importance of investigating non-ideal agents in epistemology, political philosophy, game theory, economics, and related fields. Other prominent voices in that chorus include Simon (1957), Cherniak (1981) Kahneman and Tversky (1979), and Epstein (2006).

Gutmann, A., & Thompson, D. (1996). *Democracy and disagreement*. Cambridge: Harvard University Press.
Halpern, J. Y. & Pass, R. (2010). I don't want to think about it now: Decision theory with costly computation. In *Twelfth international conference on the principles of knowledge representation and reasoning*.
Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3). http://jasss.soc.surrey.ac.uk/5/3/2.html.
Hegselmann, R., & Krause, U. (2005). Opinion dynamics driven by various ways of averaging. *Computational Economics, 25*(4), 381–405.
Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labour: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation, 9*(3), 10.
Hellman, M. A., & Cover, T. M. (1970). Learning with finite memory. *The Annals of Mathematical Statistics, 41*(3), 765–782.
Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review, 121*(2), 206–224.
Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65,* 575–603.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society, 47*(2), 263–291.
Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy, CV, 10,* 611–633.
Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5–22.
Knight, J., & Johnson, J. (2011). *The priority of democracy: Political consequences of pragmatism*. Princeton: Princeton University Press.
Landemore, H. (2013). *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton: Princeton University Press.
Lehrer, K. (1990). *Theory of knowledge*. Boulder, CO: Westview.
Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin, 18,* 669–679.
Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109.
Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making a meta-analysis. *Personality and Social Psychology Review, 16*(1), 54–75.
Lumet, S., & Rose, R. (1957). *Twelve angry men*. Los Angeles: Orion-Nova Twelve Angry Men.
McCain, K. (2014). *Evidentialism and epistemic justification*. London: Routledge.
McHoskey, J. W. (1995). Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology, 17,* 395–409. https://doi.org/10.1207/s15324834basp1703_7.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin, 23,* 636–653.
Murphy, P. (2016). Coherentism in epistemology. *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/coherent/.
Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology, 21,* 1058–1082.
Prior, M. (2013). Media and political polarization. *Annual Review of Political Science, 16,* 101–127.
Ramsey, P. F. (1926). Truth and probability. In H. E. Kyburg & H. E. K. Smokler (Eds.), *Studies in subjective probability*. Huntington, NY: Robert E. Kreiger Publishing Co.
Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 129–152). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511809477.010.
Schelling, T. C. (1969). Models of segregation. *The American Economic Review, 59*(2), 488–493.
Schroeder, M. (2010). *What makes reasons sufficient?* Unpublished manuscript, University of Southern California.

Schroeder, M. (2015). Knowledge is belief for sufficient (objective and subjective) reason. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology, (5)* (pp. 226–252). Oxford: Oxford University Press.

Sherman, D. K., Hogg, M. A., & Maitner, A. T. (2009). Perceived polarization: Reconciling ingroup and intergroup perceptions under uncertainty. *Group Processes and Intergroup Relations, 12,* 95–109.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.

Sosa, Ernest. (1985). Knowledge and intellectual virtue. *The Monist, 68,* 224–245.

Stasser, G. (1988). Computer simulation as a research tool: The DISCUSS model of group decision making. *Journal of Experimental Social Psychology, 24,* 393–422.

Stasser, G., & Birchmeier, Z. (2003). Group creativity and collective choice. In P. B. Paulus & B. A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 85–109). New York: Oxford University Press.

Strevens, M. (2003). The role of the priority rule in science. *The Journal of Philosophy, 100*(2), 55–79.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy, 10,* 175–195.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton: Princeton University Press.

Sunstein, C. R. (2017). *#Republic*. Princeton: Princeton University Press.

Sunstein, C. R., The law of group polarization (1999). University of Chicago Law School, John M. Olin Law & Economics Working Paper No. 91. Available at SSRN: https://ssrn.com/abstract=199668.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior, 31,* 137–155. https://doi.org/10.1007/s11109-008-9075-8.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755–769.

Talbott, W. (2016). Bayesian epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian.

Vitz, R. (n.d.). Doxastic volunteerism. *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002. http://www.iep.utm.edu/doxa-vol/.

Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica, 82*(6), 2257–2294.

Zollman, K. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 574–587.

Zollman, K. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.